Impact of selection bias on the evaluation of clusters of chemical compounds in the drug discovery process

Ariel Alonso[1] * **Elasma Milanzi**[2] **Geert Molenberghs**[2,3]
Christophe Buyck[4] **Luc Bijnens**[4]

[1] *Department of Methodology and Statistics. Maastricht University. The Netherlands*

[2] *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[3] *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[4] *Janssen Pharmaceutica. B-2340 Beerse, Belgium*

Abstract

Expert opinion plays an important role when selecting promising clusters of chemical compounds in the drug discovery process. Indeed, experts can qualitatively assess the potential of each cluster and, with appropriate statistical methods, these qualitative assessments can be quantified into a success probability for each of them. However, one crucial element often overlooked is the procedure by which the clusters are assigned/selected to/by the experts for evaluation. In the present work, the impact is studied that such a procedure may have on the statistical analysis and the entire evaluation process. It has been shown that some implementations of the selection procedure may seriously compromise the validity of the evaluation and, consequently, the fully random allocation of the clusters to the experts is strongly advocated.

keywords: Drug discovery, Missing data, Sensitivity analysis, Hierarchical models.

1 Introduction

Over the last decades, as a result of advances in fields like genetics and molecular biology, our capacity to develop chemical compounds for therapeutic use has been dramatically

*Corresponding author: Ariel Alonso, Department of Methodology and Statistics. Maastricht University. The Netherlands.
ariel.alonso@maastrichtuniversity.nl

increased. Nevertheless, developing these compounds into an effective drug is an expensive and lengthy process and, therefore, pharmaceutical companies need to carefully evaluate their potential before investing more resources on them (Alonso *et al.*, 2008). Nowadays, expert opinion is widely acknowledged as a crucial element in this evaluation process (Oxman *et al.*, 2007; Hack *et al.*, 2011). For this purpose, in practice, similar compounds are grouped into clusters whose potential is qualitatively assessed by experts. Further, with appropriate statistical methods, these assessments can be quantified into a success probability for each cluster, where success refers to recommending the inclusion of a cluster into the sponsor's database for future scrutiny.

The large number of clusters typically involved in these studies implies that a selection procedure, by which every expert chooses or gets assigned a number of clusters for evaluation, needs to be implemented. In the present work we argue that some implementations of the selection procedure may lead to serious selection bias that can jeopardize the entire evaluation process. Two possible strategies to avoid the previous problem are: (i) to compel every expert to evaluate all clusters and (ii) to assign a subset of the clusters to each expert fully randomly. Strategy (i) may be practically infeasible, given the exorbitant number of clusters one frequently is confronted with in this type of studies. Implementing strategy (ii) may lead to some logistic difficulties, but it arguably is the most reasonable and reliable option to avoid bias and simplify the posterior analysis of the data. In the present work we strongly advocate (ii).

Problems that come with selection bias, as well as their possible correction, have been documented in many fields (Horwitz and Feinstein, 1978; Hernán, Hernández-Diaz, and Robins, 2004; Geneletti, Richardson, and Best, 2009). Geneletti *et al.* (2011) noted that the crucial factor to determine the most appropriate bias correction method is the underlying cause of bias. This is apparent in the methods available in the literature, given that most of them are tailored towards a specific form of bias origin. For instance, Torner *et al.* (2010) addressed this issue in cohort studies, where bias may result from over-selecting severely ill patients, due to the long time taken by the less ill individuals to portray the symptoms. They tackled the problem by introducing a time window between entry into the cohort and entry into the study. Heckman (1979) approached the topic in applied econometrics using a correction, based on a two-stage estimation method, that is easy to implement and has a firm basis in statistical theory (Puhani, 2000). Selection bias may also emanate, among other sources, from missing data (nonresponse bias, attrition bias), censoring or publication bias; research efforts have been undertaken in these scenarios as well (Lee and Marsh, 2000; Baser *et al.*, 2003; Jüni and Egger, 2005).

A key similarity between many of the methods discussed by these authors is the formulation of separate models for the outcome and the selection process. Typically, untestable assumptions are associated with these models, simply because the outcomes of subjects that were not selected are never known.

In spite of earlier research efforts, the impact of selection bias has often been underestimated, as many share the school of thought that it may not be serious enough to motivate the use of complicated bias correction methods. Contrary to this we found that, in the scenario studied in this manuscript, the selection process may need to be explicitly modeled even if selection bias is not present. In addition, we showed using theoretical elements and simulations that, in the presence of selection bias, the probability of success for every cluster can be estimated only by making strong and untestable

assumptions. However, an upper bound for this probability may be obtained under a weaker condition of monotonicity.

The paper is organized as follows. A case study is introduced and analyzed in Section 2. The selection bias problem is studied in detail in Section 3 and a simulation study is presented and discussed in Section 4. Finally, the case study is reanalyzed in Section 5 and some concluding remarks are offered in Section 6.

2 The case study

The pharmaceutical company Johnson & Johnson carried out a project to assess the potential of $22,015$ clusters of chemical compounds, in order to identify those that warrant further screening. In total 147 experts took part in the study and their assessments were coded as 1 if they recommended the cluster for further screening, -1 if not recommended and 0 if indifferent. For the sake of our discussion the response was dichotomized. We adopted a coding scheme where 1 corresponds to a positive recommendation and 0 otherwise. However, the methodology presented can easily accommodate other coding schemes as well.

Experts carried out the evaluation of the clusters using the desk-top application Third Dimension Explorer (3DX) (Agrafiotis *et al*, 2007). For every expert, in a typical session, a random subset of clusters selected from the entire set of $22,015$ was assigned for evaluation. Each cluster was presented with additional information that included its size, the structure of some of its distinctive members like the compound with the lowest/highest molecular weight, and 1–3 other randomly chosen members of the cluster. The application was designed to support multiple sessions that would allow the expert to stop and resume the evaluation at their own convenience. The clusters in the subset could be evaluated in any order by the expert, but a new random subset of clusters, excluding the ones already rated, was assigned for evaluation only when all the clusters in the previous subset were evaluated, or when the expert resumed the evaluation after interrupting the previous session for a break. Clusters assigned but not evaluated could, in principle, be assigned again in another session.

The histogram in Figure 1 displays the distribution of the number of clusters evaluated by the experts. Clearly, the distribution is positively skewed, indicating that, as one would expect, many experts opted to evaluate few clusters. Indeed, 25% of the experts evaluated less than 345 clusters, 50% less than 1200 and 75% of the experts evaluated less than 2370 clusters. Evidently, the large differences in the number of clusters evaluated by the experts are not the result of the random allocation, but rather are dictated by the number of evaluation sessions each expert found convenient. Actually, the possibility of interrupting and reassuming the evaluation session at will allowed the experts to influence the selection process and, hence, standard models that assume complete randomization may not be appropriate.

2.1 Estimating the probability of success

The main goal of the study was to find out, using these qualitative evaluations, which clusters had the highest probability of being recommended for further screening and

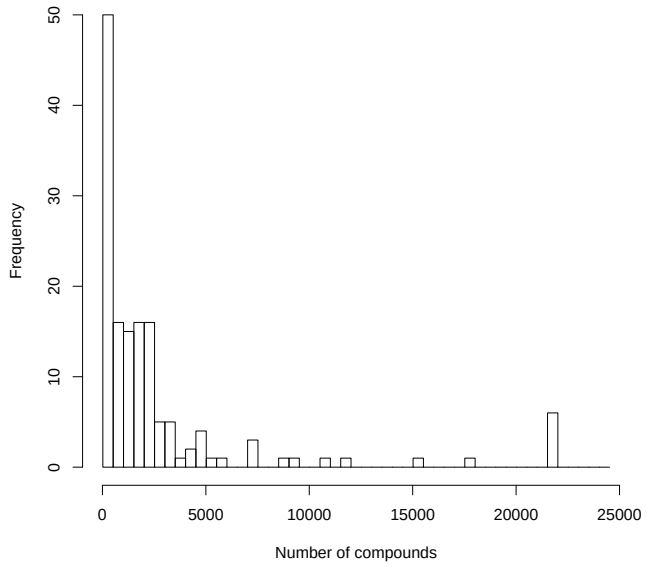

Figure 1: *Histogram for the number of clusters rated by the experts. The height of a bar indicates the number of experts whose number of rated clusters fall within the range given by the width of the bar.*

therefore, should be included into the sponsor's database. To estimate this probability of success for every cluster, let us denote the vector of ratings associated with expert i by $\boldsymbol{Y}_i = (Y_{ij})_{j \in \Lambda_i}$, where Λ_i is the subset of all clusters evaluated by the ith expert and $i = 1, \ldots, n$. Milanzi *et al.* (2013) considered the following logistic-normal model

$$\text{logit}\left[P\left(Y_{ij} = 1 | b_i\right)\right] = \beta_j + b_i, \tag{1}$$

where β_j is a fixed parameter characterizing the effect of cluster C_j with $j \in \Lambda_i$ and $b_i \sim N(0, \sigma_b^2)$ is a random expert effect. Based on model (1) these authors calculate the marginal probability of success for cluster C_j by integrating over the random effect, i.e,

$$P\left(Y_j = 1\right) = \int \frac{\exp\left(\beta_j + b\right)}{1 + \exp\left(\beta_j + b\right)} \phi(b|0, \sigma_b^2)\, db, \tag{2}$$

where $\phi(b|0, \sigma_b^2)$ denotes a normal density with mean zero and variance σ_b^2.

The likelihood emanating from model (1) suffers from a severe dimensionality problem. Indeed, the vector of fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)'$ has dimension $N = 22,015$ and the dimension (N_i) of the response vector $\boldsymbol{Y}_i$ ranges from 20 to $22,015$. As a consequence, serious computational issues can emerge when fitting model (1) with the most commonly available computing resources. Milanzi *et al.* (2013) developed an algorithm that allows to handle these issues with a very small loss of efficiency and, hence, in the present work the dimensionality problem will not be discussed further.

Results for the top 20 ranked clusters, i.e., the clusters with the highest estimated probability of success are given in the first part of Table 1 (under the 'Naive' columns).

The median estimated probability of success for all clusters was around 18%, rather a low value, and 75% of the clusters had estimated probabilities of success smaller than 29%. However, at the top 20, all clusters had an estimated probability larger than 60% and those in the top 3 had probabilities of success around 90%.

In addition, we also found a lot of heterogeneity between experts with an estimated variance $\widehat{\sigma}_b^2 \approx 20$. Note that, on the one hand, this large variance may indicate the need for selecting experts from a more uniform population by defining, for example, more stringent selection criteria. On the other hand, more stringent selection criteria may conflict with having experts that represent an appropriately broad range of expert opinion. Finding a balance between these two considerations is very important to guarantee the overall quality of the study and, in general, if substantial heterogeneity among experts is encountered, then more discussions should be held to determine the reasons for it before further actions are taken.

Finally, taking into account practical considerations like the economic cost associated with the development of these clusters of compounds, the time frame required for such a development and the social and economical gains that these clusters of compounds may bring, researchers could define the minimum probability of success that may justify the further study of a given cluster.

As previously stated, the possibility of interrupting and reassuming the evaluation sessions at will, allowed the experts to influence the selection process. This raises concerns about the possible presence of selection bias. In the next section, this important issue is studied in more detail.

3 Selection bias

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iN})'$ denote the vector containing the selection-indicators for expert i, where $X_{ij} = 1$ if expert i evaluates cluster j and 0 otherwise. The probability that expert i would rate cluster j as 1, given that he actually evaluates it, can be conceptualized as

$$P\left(Y_{ij} = 1 | X_{ij} = 1, a_i, b_i\right) = \frac{P\left(Y_{ij} = 1, X_{ij} = 1 | a_i, b_i\right)}{P\left(X_{ij} = 1 | a_i\right)}, \tag{3}$$

where $(a_i, b_i)'$ is a vector of expert-specific random effects, assumed to follow a bivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. We say that there is selection bias in the rating process if $P\left(Y_{ij} = y_{ij} | X_{ij} = 1, a_i, b_i\right) \neq P\left(Y_{ij} = y_{ij} | X_{ij} = 0, a_i, b_i\right)$. It can be easily shown that absence of selection bias is equivalent to the validity of the following conditional independence assumption

$$P\left(Y_{ij} = y_{ij}, X_{ij} = x_{ij} | a_i, b_i\right) = P\left(Y_{ij} = y_{ij} | b_i\right) P\left(X_{ij} = x_{ij} | a_i\right), \tag{4}$$

for all i, j and, consequently, in the rest of the manuscript these two concepts, lack of selection bias and conditional independence, will be used interchangeably. Essentially, (4) states that for every expert the rating and selection procedures are independent and governed by different, although possibly correlated, random effects. Some important scenarios covered by (4) are the ones described as strategy (i) and (ii) in Section 1. Indeed, in strategy (i) all experts are compelled to evaluate all clusters and, therefore,

$P(X_{ij}=1|a_i) = 1$ for all i, j. Moreover, in strategy (ii) the possible dependence between Y_{ij} and X_{ij} is broken by the random allocation and in that case typically $P(X_{ij}=1|a_i) = P(X_{ij}=1)$. Under (4), expression (3) can be rewritten as

$$P(Y_{ij}=1|X_{ij}=1,a_i,b_i) = P(Y_{ij}=1|X_{ij}=1,b_i) = P(Y_{ij}=1|b_i). \tag{5}$$

Model (1), used in Section 2 to quantify the success probabilities, basically tries to characterize $P(Y_{ij}=1|b_i)$ and, hence, it is valid if the conditional independence assumption holds. Some comments are in place. Note first that, on the one hand, $P(Y_{ij}=1|b_i)$ quantifies the chance that expert i will rate cluster j as 1, irrespective of whether he actually evaluates the cluster or not. Thus, it is a marginal probability that does not depend on the selection process. On the other hand, $P(Y_{ij}=1|X_{ij}=1,b_i)$ describes the chance that expert i will rate cluster j as 1 given that he evaluates it and, in principle, it might differ from $P(Y_{ij}=1|X_{ij}=0,b_i)$. Actually, in the most general scenario, the potential of cluster j can be quantified as

$$P(Y_j=1) = \int\int P(Y_{ij}=1|a_i,b_i)\ \phi(a_i,b_i|\mathbf{0},\mathbf{\Sigma})\, da_i db_i, \tag{6}$$

where $\phi(\cdot|\mathbf{0},\mathbf{\Sigma})$ denotes a bivariate normal density with mean zero and covariance matrix $\mathbf{\Sigma}$ and

$$\begin{aligned} &P(Y_{ij}=1|a_i,b_i) = E_X\left[P(Y_{ij}=1|X_{ij}=x_{ij},a_i,b_i)\right] \qquad (7)\\ &= P(Y_{ij}=1|X_{ij}=1,a_i,b_i)\,P(X_{ij}=1|a_i) + P(Y_{ij}=1|X_{ij}=0,a_i,b_i)\,P(X_{ij}=0|a_i). \end{aligned}$$

This expression is very insightful. Note first that we have information about how the experts rated the clusters they evaluated and, therefore, $P(Y_{ij}=1|X_{ij}=1,a_i,b_i)$ can be estimated from the data. We also have information about which clusters every expert evaluated and we could use this information to estimate $P(X_{ij}=1|a_i)$. The critical term in (7) is $P(Y_{ij}=1|X_{ij}=0,a_i,b_i)$. In fact, the event $\{Y_{ij}=y_{ij}|X_{ij}=0,a_i,b_i\}$ is counterfactual and we do not have information about how the experts would have rated a cluster they did not evaluate if, contrary to fact, they had evaluated it. As a result, this probability is not identifiable from the data without additional assumptions.

The previous discussion illustrates that in the most general case computing (6) requires: (1) to explicitly model $P(X_{ij}=1|a_i)$ and (2) to make untestable assumptions about the *counterfactual* probabilities $P(Y_{ij}=1|X_{ij}=0,a_i,b_i)$. A reasonable such assumption in many situations may be the following monotonicity condition

$$P(Y_{ij}=1|X_{ij}=0,a_i,b_i) \leq P(Y_{ij}=1|X_{ij}=1,a_i,b_i).$$

That may be the case, for instance, if experts choose to evaluate those clusters they find more promising or interesting. The previous inequality implies that $P(Y_{ij}=1|a_i,b_i) \leq P(Y_{ij}=1|X_{ij}=1,a_i,b_i)$ and, hence, one could use the data at hand to provide an upper bound for (6). This upper bound suggests that in many applications discarding those clusters with a small estimated probability of success may be reasonable, even if selection bias is present. Nonetheless, one should be cautious when interpreting a large probability of success if selection bias is suspected.

3.1 How ignorable is the selection procedure in the absence of selection bias?

It is clear from the previous discussion that, in the presence of selection bias, one needs to explicitly model the selection mechanism to compute (6). Nevertheless, the preceding arguments do not fully clarify whether the selection procedure can be safely ignored when selection bias is not present. In what follows we will study the ignorability of the selection process in absence of selection bias, but first we need to extend notation. Let $P\left(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j\right)$ and $P\left(Y_{ij} = y_{ij}|X_{ij} = x_{ij}, b_i, \beta_j\right)$ denote the models for the selection and rating procedure respectively. Note that in the previous formulation we allow the selection procedure to depend on the parameters that characterize the rating process (β_j) and also on other selection-specific parameters (α_j). Furthermore, it will be assumed that the components of the vectors $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are independent conditionally on the random effects a_i and b_i, respectively. It is easy to see that, in absence of selection bias, (6) takes the simpler form

$$P\left(Y_j = 1|\beta_j\right) = \int P\left(Y_{ij} = 1|X_{ij} = 1, b_i, \beta_j\right)\ \phi(b_i|0, \sigma_b^2)\, db_i. \tag{8}$$

Expression (8) does not depend on the selection procedure and the estimation of the success probabilities is reduced to the estimation of the clusters effect and the variance component σ_b^2. However, even though the selection procedure does not explicitly appear in (8), one may need to take it into account when estimating the β_js and σ_b^2.

In fact, one estimates these parameters using the complete data $\boldsymbol{Y}_i, \boldsymbol{X}_i \in \{0,1\}^N$. The vector of ratings can be decomposed as $\boldsymbol{Y}_i = (\boldsymbol{Y}_{0i}', \boldsymbol{Y}_{1i}')'$, where $\boldsymbol{Y}_{1i} \in \{0,1\}^{N_i}$ is the sub-vector associated with the clusters the expert evaluated, $\boldsymbol{Y}_{0i}$ is the obvious complement and $N_i = \mathbf{1}'\boldsymbol{X}_i$. The joint distribution of $(\boldsymbol{Y}_i', \boldsymbol{X}_i', a_i, b_i)'$ takes the form

$$P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{X}_i = \boldsymbol{x}_i, a_i, b_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right) = P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i, b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}).$$

Under conditional independence $P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i, b_i, \boldsymbol{\beta}\right) = P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\beta}\right)$. One further has that $P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\beta}\right) = P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{Y}_{0i} = \boldsymbol{y}_{0i}|b_i, \boldsymbol{\beta}\right)$ and, therefore,

$$\begin{aligned}
&P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i, a_i, b_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right) \\
&= \sum_{\boldsymbol{y}_{0i}} P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}), \\
&= \sum_{\boldsymbol{y}_{0i}} P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{Y}_{0i} = \boldsymbol{y}_{0i}|b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}), \\
&= P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}), \\
&= \left[\prod_{j \in \Lambda_i} P\left(Y_{1ij} = y_{1ij}|b_i, \beta_j\right)\right] \left[\prod_{j}^{N} P\left(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j\right)\right] \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}).
\end{aligned}$$

Marginally, the previous equations lead to

$$P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right) = \int\int P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\ \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma})\, da_i db_i, \tag{9}$$

and the likelihood emerging from (9) takes the form

$$L\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right) = \prod_{i}^{n} P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_{i} = \boldsymbol{x}_{i} | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right). \tag{10}$$

Using the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\boldsymbol{\alpha}}_n$, $\widehat{\sigma}^2_{bn}$ one can estimate the probabilities of success by substituting $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\sigma}^2_{bn}$ into (8). Note, however, that to estimate $\boldsymbol{\beta}$, σ^2_b, one may need to explicitly model the selection process. An important special instance where the selection mechanism can be ignored is when the selection and rating processes are also marginally independent, i.e, when $\phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}) = \phi(a_i|0, \sigma^2_a)\phi(b_i|0, \sigma^2_b)$ and have a disjoint parametric space. In fact, under these assumptions (9) simplifies to

$$P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_{i} = \boldsymbol{x}_{i} | \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2\right) = \int P\left(\boldsymbol{X}_{i} = \boldsymbol{x}_{i} | a_i, \boldsymbol{\alpha}\right)\, \phi(a_i|0, \sigma^2_a)\, da_i \int P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i} | b_i, \boldsymbol{\beta}\right)\, \phi(b_i|0, \sigma^2_b)\, db_i.$$

Consequently, regarding the parameters of interest $\boldsymbol{\beta}$ and σ^2_b, the contribution of expert i to the likelihood becomes

$$\int P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i} | b_i, \boldsymbol{\beta}\right)\, \phi(b_i|0, \sigma^2_b)\, db_i, = \int \left[\prod_{j \in \Lambda_i} P\left(Y_{1ij} = y_{1ij} | b_i, \beta_j\right)\right]\, \phi(b_i|0, \sigma^2_b)\, db_i.$$

The previous expression is the contribution of expert i to the likelihood when the selection mechanism has been discarded. Therefore, in this scenario, if conditional independence holds, the selection procedure can be fully ignored.

Importantly, such a scenario will result if a random allocation of the clusters to experts is implemented, where the experts have not influence whatsoever on the selection process. The previous discussion shows that fully random allocation is a powerful tool not only to avoid selection bias, by guaranteeing conditional independence, but also to considerably simplify the analysis by making the selection mechanism ignorable for the estimation of the parameters.

4 Simulation study

To numerically evaluate the ignorability of the selection procedure and the impact of selection bias on the assessments, a simulation study was designed. The data were generated mimicking the main characteristics encountered in the case study. Nonetheless, the size of the simulated data sets were chosen so that model (1) could be fitted using maximum likelihood. To that effect, two hundred data sets were generated, with the following parameters held constant across data sets: (1) Number of clusters $N = 50$, chosen to ensure tractability of maximum likelihood estimation for the whole data, (2) number of experts $n = 147$, and (3) a set of 50 values assigned to the parameters characterizing the cluster effects (β_j), which were sampled from a $N(0, 2)$ one time and then held constant in all data sets. Factors varying across the data sets were: (1) the number of ratings per expert N_i, determined by the selection model and (2) a set of 147 expert random-effects b_i, independently sampled from $N(0, 10)$. Conceptually, each generated data set represents a replication of the evaluation study in which a new set of experts rates the same clusters. Therefore, varying b_i from one data set to another

resembles the use of different groups of experts in each study, sampled from the entire experts' population. Furthermore, the selection and rating probabilities were computed using the following models

$$\text{logit}\left[P\left(X_{ij}=1|b_i\right)\right]=b_i, \tag{11}$$
$$\text{logit}\left[P(Y_{ij}=1|b_i)\right]=\beta_j+b_i, \tag{12}$$

and $X_{ij}\sim$ Bernoulli $\left[P\left(X_{ij}=1|b_i\right)\right]$, $Y_{ij}\sim$ Bernoulli $\left[P\left(Y_{ij}=1|b_i\right)\right]$, respectively. Models (11) and (12) are a special case of the general modeling framework introduced in Section 3.1. In fact, to simplify the computational burden and improve numerical stability, we considered the situation in which the selection and rating procedures shared a common random effect. This is the so-called shared parameter model (SPM), for which corr$(a_i,b_i)=1$ (Follmann and Wu, 1995; Little, 1995).

In the previous setting, like in the case study, some experts will tend to evaluate a large number of clusters whereas others will tend to evaluate only a reduced number of them. Note further that the rating process does not depend on the selection procedure, i.e.,

$$P\left(Y_{ij}=1|X_{ij}=1,b_i,\beta_j\right)=P\left(Y_{ij}=1|X_{ij}=0,b_i,\beta_j\right)=P\left(Y_{ij}=1|b_i\right),$$

and, therefore, there is no selection bias. All the generated data sets were analyzed using model (1) and the success probability of each cluster was estimated by plugging the necessary maximum likelihood estimators into (2). The integral was approximated as

$$P_{S0}=P\left(Y_j=1\right)=\sum_{q=1}^{Q}\frac{\exp\left(\beta_j+b_q\right)}{1+\exp\left(\beta_j+b_q\right)}$$

where $Q=10,000$ and $b_q\sim N(0,\widehat{\sigma}_b^2)$ when using the $\widehat{\beta}_j$ values estimated from model (1) and $b_q\sim N(0,10)$ when using the true β_j values. Table 2 summarizes the main results and the clusters are ordered decreasingly according to their true probability of success. Clearly, ignoring the selection procedure can have a huge impact on the estimators $\widehat{\beta}_j$ and, consequently, on the estimates of the success probabilities. Indeed, using the estimated probability of success $\widehat{P}_{S0}$, cluster 32 would be considered the most promising one whereas, in reality, it should be ranked as number 8 taking into account its true probability of success. These findings unequivocally showed that ignoring the selection process, when estimating the model parameters and the probabilities of success, may be extremely misleading even in the absence of selection bias.

Further, we studied a scenario in which selection bias was present. To this end we considered the following rating mechanism

$$\text{logit}\left[P(Y_{ij}=1|X_{ij}=x_{ij},b_i)\right]=\begin{cases}\beta_j+b_i & \text{if } x_{ij}=1,\\ \beta_j+b_i-0.223 & \text{if } x_{ij}=0.\end{cases} \tag{13}$$

Essentially, (13) implies that, for every expert i, the odds of rating a cluster as 1 is 25% larger when the cluster is evaluated than when it is not. The values of the true success probabilities in this scenario, computed using (7), are given under the column P_{S1} in Table 2. Note that, even if one can avoid bias when estimating β_j and σ_b^2, a comparison

between P_{S0} and P_{S1} clearly shows that, in the presence of selection bias, a naive use of (2) would lead to an overestimation of the true probabilities of success, as it was stated in Section 3.1.

In a second simulation study the selection process was taken into account when estimating the parameters of interest. Basically, the likelihood (10) was maximized with the selection proccess modeled as logit $[P(X_{ij} = 1|a_i)] = \alpha + a_i$. The settings were essentially the same as before but, to alleviate the computational burden, only 10 clusters were considered. The results are presented in Table 3. Once more, the naive approach that ignores the selection process led to biased estimates for the cluster effects, the variance component and the probabilities of success. Importantly, for some clusters, the relative bias in the estimated probability of success was as large as 30%. Further, when the selection procedure was incorporated into the likelihood as given in (10), the bias disappeared and the probabilities of success were always accurately estimated. Additional simulations (not shown) with a reduced number of 50 experts confirmed these conclusions.

5 Case study revisited

In a new analysis of the case study, the parameters of interest were estimated using likelihood (10) with the selection process modeled as logit $[P(X_{ij} = 1|a_i)] = \alpha_j + a_i$. The main results are presented in the second part of Table 1 (under the 'Joint Model' columns). Both modeling approaches 'Naive' and 'Joint' assume absence of selection bias, but while the naive-model fully ignores the selection process, the joint-model does take the selection process into account when estimating the relevant parameters.

There are substantial differences between the results obtained with both methods. In general, the joint-model approach seems to produce lower estimates of the success probabilities and leads to a different ranking of the clusters. For instance, cluster 432169 ranked as number one by the joint-model approach with an estimated probability of success 0.91, was ranked as 18 by the naive-model method with an estimated probability of success 0.64. Additionally, the joint-model also produced a smaller estimate of the between-raters variability.

For completeness, the third part of Table 1 (under the 'SPM' columns) shows the results obtained with the shared parameter model introduced in Section 4. The differences between the SPM and the other two methods are striking. In fact, the SPM produces much lower estimates of the success probabilities and, therefore, it provides a rather sceptical view of the potential of all clusters. In addition, it also produces a much smaller estimate of the between-experts variability. However, there is some evidence to suggest that the SPM may not be a good description of data generating mechanism. Indeed, the shared parameter model makes some testable predictions that allow to evaluate its adequacy. For instance, it postulates that experts who rate more clusters should tend to give higher ratings as well. Figure 2 shows that this prediction of the model is not fulfilled by the data at hand and raises doubts about its adequacy.

Arguably, the joint-model approach offers a more flexible description of reality and, therefore, one may be inclined to put more weight on the results emanating from it. Nonetheless, it is important to point out that a formal model comparison between candidate models, based on maximum likelihood tests or information criteria, is hampered

in this scenario by the fact that the models are not fitted using maximum likelihood. One can, however, evaluate the fit produced by the different selection models using the data. Given that the rating model is the same in all cases, one could use the model fit of the selection model as an informal criterion to select the best joint model. In the case study, the conclusions emanating from the joint model get additional support from the fact that the selection model it uses, offers the most plausible description of the data. In any case, a very careful discussion incorporating domain-specific knowledge, will be needed before final conclusions can be drawn from these analyses.

6 Discussion and concluding remarks

The topic studied here can be related to other statistical fields and perhaps the most evident connection is with missing data analysis. Indeed, like many problems from areas like hierarchical models (Lindstrom and Bates, 1988), causal inference, and treatment compliance (Holland, 1986), selection bias could also be framed within a missing data context. To illustrate this connection using a simpler notation, let us focus on the special case in which the selection and rating procedures shared a common random effect. Conditioning on the expert effect, one could think of the selection and rating procedures introduced in Section 3, as analogous to the pattern mixture framework often use to handle missing observations (Molenberghs and Kenward, 2007). Similarly, the condition used to define selection bias in Section 3, is closely related to the concept of *missing not at random* (MNAR) that appears in the classical missing data taxonomy (Rubin, 1976; Kenward and Carpenter, 2007; Molenberghs and Kenward, 2007), and which means that the missing-data mechanism is related to unobserved outcomes, in addition to observed outcomes and covariates. To exemplify this, consider the expression

$$P\left(Y_{ij}=y_{ij}|X_{ij}=x_{ij},b_i\right)=P\left(X_{ij}=x_{ij}|Y_{ij}=y_{ij},b_i\right)\frac{P\left(Y_{ij}=y_{ij}|b_i\right)}{P\left(X_{ij}=x_{ij}|b_i\right)}. \qquad (14)$$

If the probability of not evaluating a cluster is independent of its (unobserved) rating, then we have $P\left(X_{ij}=0|Y_{ij}=y_{ij},b_i\right)=P\left(X_{ij}=0|b_i\right)$, which is the definition of the *Missing At Random* mechanism (MAR) in the Rubin taxonomy (Rubin, 1976). MAR means that, given observed outcomes and covariates, missingness does not further depend on unobserved ones. It is easy to see that (14) and the subsequent expressions imply

$$P\left(Y_{ij}=y_{ij}|X_{ij}=1,b_i\right)=P\left(Y_{ij}=y_{ij}|X_{ij}=0,b_i\right)=P\left(Y_{ij}=y_{ij}|b_i\right),$$

and, therefore, the absence of selection bias can be seen as an MAR process, given the expert. Moreover, the conditional independence assumption for the rating and selection procedure introduced in Section 3, is closely related to the generalized shared parameter modeling (GSPM) framework, used to describe a MNAR mechanism (Creemers *et al.*, 2011). This relationship with the GSPM explains why, unlike in the selection model context in missing data, where under MAR the likelihood paradigm implies ignorability, in the context studied in this manuscript even in absence of selection bias the selection procedure will often be non-ignorable. The reason for this important difference is that the random effects governing the selection and rating procedures are correlated and, therefore, marginally independence does not hold.

It has been shown that in a missing data problem the data at hand do not provide enough information to discriminate between MAR and MNAR (Molenberghs *et al.*, 2008). Likewise, the data at hand will not provide enough information to discard the presence of selection bias if the assignment mechanism was non-random or had the potential to be influenced by the experts. One could, however, conceive a sensitivity analysis to evaluate the robustness of the conclusions with respect to the potential presence of selection bias.

In addition, the relevance of the numerical procedures should not be overlooked when working with complicated hierarchical models. For instance, given the complexity of the models used in the analyses of the case study and the high dimensionality of the data, in all the approaches shown Table 1 the likelihood was computed using the Laplace approximation. Unlike in the case study, the data used in the simulations had purportedly a lower dimension and this allowed to approximate the likelihood using adaptive Gaussian quadrature. It has been shown that these type of choices may have a non-negligible impact on the results (Lesaffre and Spiessens, 2001). More complex models are often less biased, but they may require a cruder approximation of the likelihood. Simpler models often allow a better approximation of the likelihood, but they may also be more prone to serious bias. The optimal balance between complexity and precision is difficult to determine in real examples where the true is unknown and this difficulty emphasizes the importance of using all information available when interpreting the results in the decision making process.

Summarizing, we have shown that the mechanism used to assign the clusters to the experts is a key issue to guarantee the validity of the entire evaluation process. Essentially, to guarantee this validity, one needs to ensure that the selection and rating processes are independent. In addition, to be able to carry out a simpler analysis that is consequently less prone to error, one also needs to ensure that both processes are marginally independent and do not share any parameter. A fully random allocation of the clusters to the experts seems to be the most, if not the only, practical way to achieve these conditions. Therefore, we strongly advocate for its use in the present work.

In cases where a fully random allocation is not feasible due to practical problems, a joint modeling approach and/or a sensitivity analysis may be the most reasonable and sound alternatives.

Acknowledgment

Elasma Milanzi and Geert Molenberghs gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy). The authors are grateful to Johnson & Johnson for the kind permission to use their data. For the computations, simulations and data processing, we used the infrastructure of the VSC — Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government — department EWI.

References

Agrafiotis DK, Alex S, Dai H, Derkinderen A, Farnum M, Gates P, Izrailev S, Jaeger EP, Konstant P, Leung A, Lobanov VS, Marichal P, Martin D, Rassokhin DN, Shemanarev

M, Skalkin A, Stong J, Tabruyn T, Vermeiren M, Wan J, Xu XY, Yao X. Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model* 2007; **47**: 1999-2014.

Alonso A, Molenberghs G. Surrogate endpoints: Hopes and perils. *Pharmacoeconomics and Outcomes Research* 2008; **8**, 255-259. DOI:10.1586/14737167.8.3.255

Baser O, Bradley CJ, Gardiner JC, and Given C. Testing and correcting for non-random selection bias due to censoring: An application to medical costs. *Health Services & Outcomes Research Methodology* 2003; **4**:93–107.

Creemers A, Hens N, Aerts M, Molenberghs G, Verbeke G, Kenward MG. Generalized shared-parameter models and missingness at random. *Statistical Modeling* 2011;**11**: 279-311.

Diggle PJ, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* 1994; **43**: 49-93.

Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 1995; **51**: 151-168.

Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics* 2009;**10**: 17-31.

Geneletti S, Mason A, Best N. Adjusting for selection effects in epidemiologic studies: Why sensitivity analysis is the only "solution." *Commentary in Epidemiology* 2011; **22**: 36-39.

Hack MD, Rassokhin DN, Buyck C, Seierstad M, Skalkin A, ten Holte P, Jones TK, Mirzadegan T Agrafiotis DK. Library enhancement through the wisdom of crowds. *Journal of Chemical Information and Modeling* 2011; **51**: 3275-3286.

Heckman J. Sample selection bias as a specification error. *Econometrica* 1979; **47**: 153-161,

Hernán MA, Hernández-Diaz S Robins, JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615-625.

Holland, PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**: 945-960.

Horwitz R, Feinstein A. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine* 1978**299**: 368-387.

Jüni P, Egger M. Empirical evidence of attrition bias in clinical trials. *International Journal of Epidemiology* 2005; **34**: 87-88.

Lee B, Marsh LC. Sample selection bias correction for missing response observations. *Oxford Bulletin of Economics and Statistics* 2000; **62**: 305-322.

Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Appl. Statist.* 2001; 50:325-335.

Lindstrom ML, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 1988; **83**: 1014-1021.

Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**: 1112-1121.

Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 2007; **16**: 199-218.

Milanzi E, Alonso A, Buyck C, Molenberghs G, Bijnens L. A permutational-splitting sample procedure to quantify expert opinion on chemical compounds using high-dimensional data. *submitted* 2013.

Molenberghs G, Beunckens C, Sotto C, and Kenward MG. Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B* 2008; **70**: 371-388.

Molenberghs G and Kenward MG. *Missing Data in Clinical Studies*. New York: Wiley,2007.

Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. New York: Springer,2005.

Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations.*Lancet* 2007; **369**: 1883-1889.

Puhani, PA. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* 2000;**14**: 53-68.

Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581-592.

Torner A, Duberg A, Dickman P, Svensson A. A proposed method to adjust for selection bias in cohort studies. *American Journal of Epidemiology* 2010; **171**: 602-608.

Table 1: *Results for the top 20 clusters according to the naive approach that ignores the selection process and the joint and shared parameter (SPM) models that take the selection process into account.Given are the estimated cluster effect ($\widehat{\beta}$), rank assigned to the cluster according to its probability of success (R), estimated success probabilities ($\widehat{P}$) and confidence interval limits (lcl, ucl). In the joint model/SPM the selection process was modeled as* $logit\{P(X_{ij} = 1|a_i/b_i)\} = \alpha_j + a_i/\alpha_j + b_i$

	Naive					Joint					SPM				
ID	R	$\widehat{\beta}$	$\widehat{P}$	lcl	ucl	R	$\widehat{\beta}$	$\widehat{P}$	lcl	ucl	R	$\widehat{\beta}$	$\widehat{P}$	lcl	ucl
265222	1	2.52	0.94	0.78	0.98	3	2.67	0.72	0.45	0.89	15	-1.10	0.34	0.16	0.59
295061	2	3.83	0.92	0.66	0.98	4	2.61	0.71	0.48	0.87	1	-0.58	0.42	0.17	0.71
359957	3	0.49	0.87	0.72	0.94	330	-0.25	0.48	0.18	0.79	47	-1.47	0.29	0.12	0.54
69850	4	1.07	0.82	0.33	0.97	182	0.11	0.50	0.23	0.77	9	-0.91	0.37	0.16	0.65
84163	5	5.24	0.77	0.41	0.97	9	1.83	0.65	0.21	0.97	239	-2.03	0.22	0.08	0.49
296443	6	2.59	0.76	0.49	0.93	10	1.62	0.64	0.33	0.87	488	-2.31	0.19	0.09	0.36
7451	7	1.28	0.74	0.16	0.96	55	0.66	0.56	0.24	0.81	272	-2.06	0.22	0.10	0.42
277619	8	1.65	0.73	0.41	0.94	89	0.44	0.54	0.17	0.87	191	-1.97	0.23	0.09	0.47
315928	9	2.04	0.72	0.37	0.92	14	1.47	0.62	0.28	0.83	5	-0.73	0.39	0.19	0.63
296535	10	2.77	0.71	0.48	0.87	5	2.37	0.70	0.38	0.91	3	-0.65	0.40	0.17	0.69
313914	11	2.18	0.70	0.40	0.89	7	2.06	0.68	0.28	0.91	166	-1.91	0.23	0.05	0.63
277774	12	2.20	0.69	0.43	0.87	20	1.30	0.61	0.37	0.81	18	-1.11	0.34	0.16	0.59
178994	13	1.85	0.68	0.45	0.84	11	1.57	0.64	0.34	0.84	2	-0.63	0.41	0.22	0.63
296560	14	1.89	0.66	0.43	0.83	8	1.86	0.66	0.39	0.85	17	-1.07	0.34	0.16	0.59
464822	15	1.21	0.66	0.43	0.83	72	0.56	0.55	0.31	0.77	45	-1.45	0.29	0.11	0.60
265441	16	1.87	0.65	0.41	0.86	15	1.44	0.62	0.34	0.85	175	-1.95	0.23	0.08	0.55
292805	17	1.47	0.65	0.38	0.84	19	1.20	0.61	0.29	0.84	4	-0.72	0.39	0.17	0.68
432169	18	1.45	0.64	0.35	0.86	1	6.26	0.91	0.51	0.99	20	-1.09	0.34	0.10	0.70
292579	19	1.85	0.64	0.24	0.90	13	1.50	0.63	0.21	0.89	48	-1.48	0.29	0.10	0.62
278927	20	1.30	0.63	0.41	0.81	76	0.51	0.54	0.31	0.76	59	-1.61	0.27	0.12	0.52
$\widehat{\sigma}^2$		20.02					18.61					4.05			

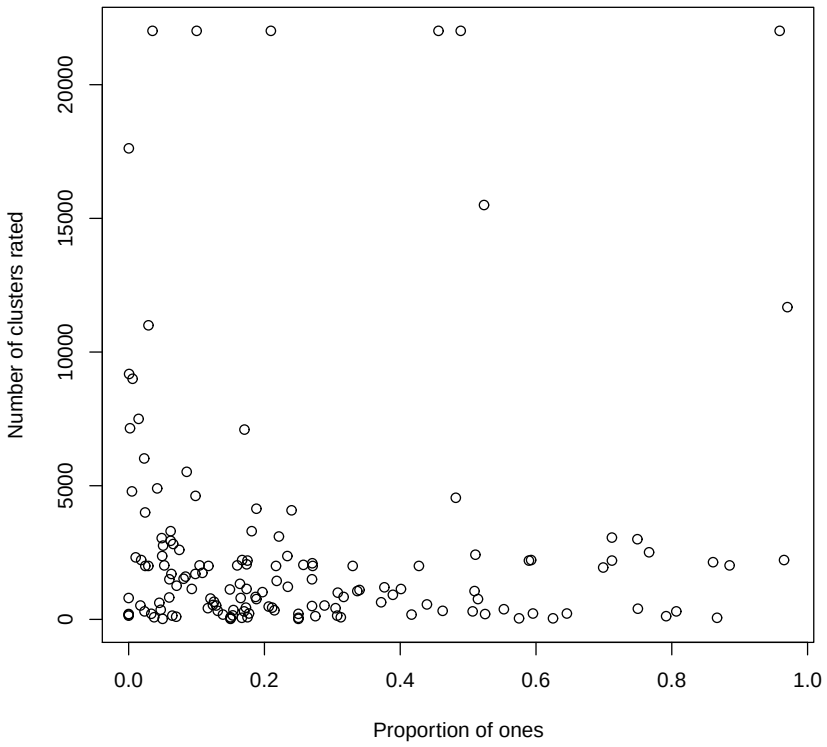

Figure 2: *A scatter plot for the number of clusters rated and the proportion of clusters recommended by each expert*

Table 2: *Simulation results. ID: cluster id; β_j: true cluster effect on the rating process; P_{S0}: true probability of success without selection bias; P_{S1}: true probability of success with selection bias. The mean of the* naive *estimated values (ignoring the selection process) are denoted using a hat$\widehat{\ }$.*

	True values			Naive	
ID	β_j	P_{S0}	P_{S1}	$\widehat{\beta}_j$	$\widehat{P}_{S0}$
3	4.326	0.865	0.858	2.338	0.746
1	3.602	0.821	0.813	-0.259	0.471
33	3.518	0.815	0.807	-0.320	0.463
47	3.434	0.809	0.801	3.146	0.808
50	3.037	0.781	0.772	1.683	0.684
27	2.127	0.706	0.696	-1.216	0.364
30	2.059	0.700	0.690	1.272	0.642
32	2.056	0.700	0.690	10.228	0.947
14	1.892	0.685	0.675	2.374	0.749
7	1.701	0.668	0.657	1.591	0.676
9	1.505	0.650	0.639	3.366	0.804
48	1.369	0.637	0.625	1.950	0.711
10	1.293	0.629	0.618	2.581	0.767
21	1.032	0.604	0.592	1.690	0.685
11	0.876	0.588	0.577	-1.637	0.320
26	0.873	0.588	0.577	4.348	0.863
15	0.685	0.569	0.558	1.671	0.683
13	0.602	0.561	0.549	4.249	0.851
4	0.582	0.559	0.547	1.827	0.698
42	0.389	0.540	0.528	1.314	0.646
σ_b^2	10.00			9.080	

Table 3: *Simulation results.* β_j*: true value used to generate the data;* P_{S0}*: true probability of success. The mean of the estimated values are denoted using the hat symbol. The estimates are obtained using the naive approach that ignores the selection process and the joint model that takes this process into account.*

	True values		Naive		Joint Model	
cid	β_j	P_{S0}	$\widehat{\beta}_j$	$\widehat{P}_{S0}$	$\widehat{\beta}_j$	$\widehat{P}_{S0}$
1	3.60	0.84	6.16	0.95	5.02	0.85
2	-1.98	0.29	-0.96	0.37	-2.01	0.29
3	4.33	0.88	9.58	0.97	7.96	0.90
4	0.58	0.56	1.57	0.70	0.59	0.56
5	0.11	0.51	1.07	0.64	0.10	0.51
6	-0.53	0.44	0.45	0.56	-0.54	0.44
7	1.70	0.68	2.75	0.82	1.73	0.68
8	-0.10	0.49	0.89	0.62	-0.08	0.49
9	1.51	0.66	2.51	0.80	1.51	0.66
10	1.29	0.64	2.29	0.78	1.31	0.64
$\widehat{\sigma}_b^2$	10.00		7.103		10.28	